(* COUNTS *)

27x

Counts of counts of species
[very long tails; rank-size-log techniques]

		Exh	Page
Introduction	27x-1		
		A	
27xA.	Counts of species 27x-1	B	
		1	27x-4
27xB.	Occurrences as basic 27x-3	2	27x-6
	- keeping the area principle 27x-7	3	27x-8
		C	
27xC.	Rank-size-log techniques 27x-9	4	27x-10
	- (sum of ranks)×(sum of sizes) 27x-11	D	
		5	27x-13
27xD.	Examples of RSL use 27x-12	6	27x-14
	- back to Josselson 27x-12	7	27x-15/16
		8	27x-17
		E	
27xE.	Raising the ends 27x-18	9	27x-19/22
		F	
27xF.	Broader usefulness 27x-23	10	27x-24
	- long-tailed counts of instances 27x-25	G	
		11	27x-26
27xG.	Further improvements 27x-25	12	27x-28
	- better scales 27x-27	13	27x-30
	- the horizontal axis 27x-29	14	27x-32
	- back to the birds 27x-29	H	
	- and for counts of species? 27x-29	15	27x-36
		P	
	**********	16	27x-39
27xH.	How far have we come? 27x-34	17	27x-40
	**********		
27xP.	Problems for chapter 27x 27x-37		

27x

Counts of counts of species
[very long tails; rank-size-log techniques]

As we tried to make clear at the beginning of chapter 27, species are different from instances, especially in terms of what "more data" means.

27xA. Counts of species

Counts of species can often be regarded as really counted fractions. If we were to get "an equal amount of data like what we have" when we have a set of counts of species we would expect the old and new data combined to have:

- about twice as large a count as before for any species for which we now have several counts -- or originally had more than a few.
- the appearance -- with one or two counts each -- of certain species not represented in the original data.

Subject always to the accidents of collection and the evils of selection, we usually find it easier to INTERPRET counts of species as counted fractions or counted rates -- in such terms as

$$\frac{\text{count for this species}}{\text{total count for all species}}$$

(a counted fractions), or

$$\frac{\text{count for this species}}{\text{"size" of data}}$$

(a counted rate), where "size" can be measured in such diverse units as area covered, time covered, effort used in search, volumes read through, and the like. This is particularly time when there are only a few species which are named, so that we can pay attention to individual counts. In such cases we are not likely to concentrate on counts of the basic counts. The methods of the next two chapters -- chapters 28 and 29 -- are likely to be appropriate.

When we have many species, so that we are likely to be forced to look at counts of basic counts, the situation is likely to be rather different. The only natural first step may be to look at counts of basic counts. The pattern of distribution of the basic counts may be very j-shaped indeed, ranging from many ones and twos up through moderate counts to a few very large counts. Ordinary histograms and rootograms just look very j-shaped. Using roots or logs for the base is likely to weaken this to just j-shaped, but we still find it hard to look at single sets of data or to make effective comparisons among two or more sets. We need a new approach.

27xB. Occurrences as basic

There is an important sense in which the occurrence replaces the thing counted -- here the species -- as the target of our attention. As a result we ought to try to use the area principle in a different form, namely:

equal area for each occurrence

This calls for relatively very large areas for very common species and is likely to make us want to make substantial changes in how we handle and exhibit our data.

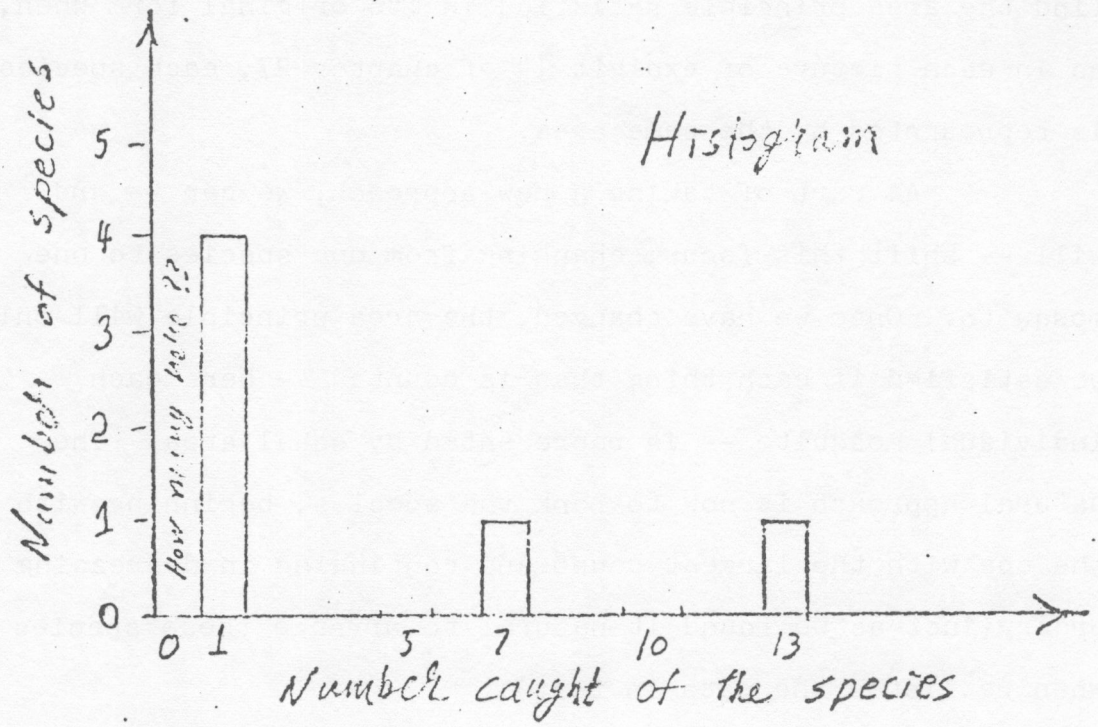
A small example can help us to see both problems and solutions. In 1951, Lumsden reported on the number of various kinds of mosquitoes caught at Tokwe, Bwamba, Uganda during a five-day period (9th-14th July 1948). One catching station was described as "forest canopy" and at it six species of *Aedes* were taken (in addition to various species of *Taeniorhynchus* and *Culex*). The numbers caught were:

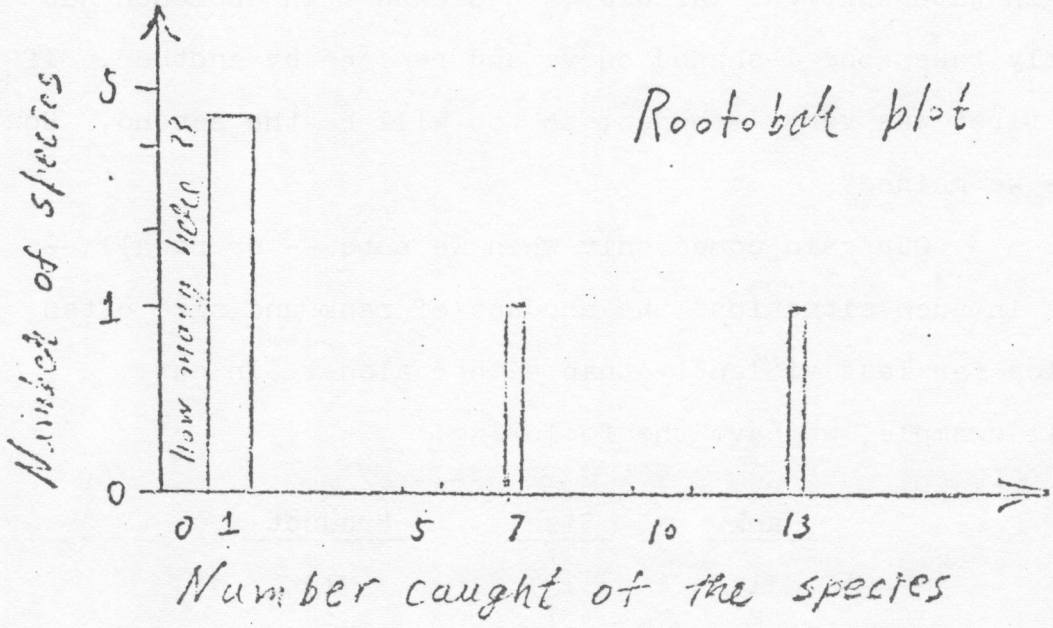
A.	*longipalpis*	13
A.	*africanus*	7
A.	*nigerrimes*	1
A.	*grahami*	1
A.	*simpsoni*	1
A.	*apicoargenteus*	1

The usual histogram or rootobar plot would appear as in exhibit 1. In it the focus of attention is the species. We

exhibit 1 of chapter 27x -- Aedes catches

Conventional plots for catches of Aedes mosquitoes





Source: W. H. R. Lumsden 1951. Probable insect vectors of yellow fever virus, from monkey to man, in Bwamba County, Uganda. Bull. Entomolog. Res. 42: 317-330. (Spec. first column of Table III at page 323.)

find the area principle satisfied in its original form when, as in each picture of exhibit 14 of chapter 27, each species is represented by the same area.

As part of taking a new approach, we can -- and will -- shift this focus, changing from one species to one mosquito. Once we have changed, the area principle will only be satisfied if each thing that is counted -- here each individual mosquito -- is represented by equal area. The natural approach is now to rank the species, beginning with the one with the largest count and continuing in decreasing order (just as we found it natural to arrange these species when we listed the data in text).

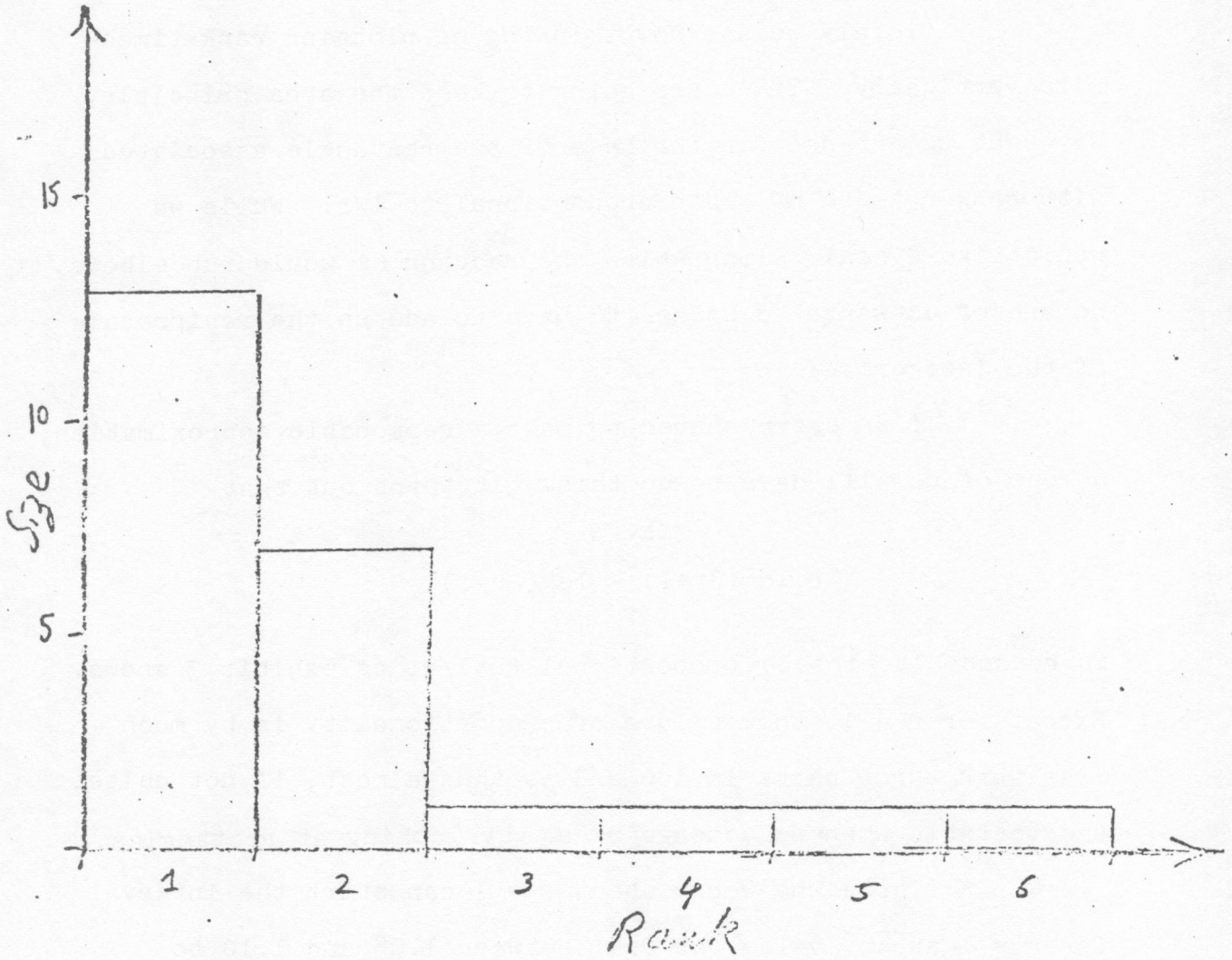
Exhibit 2 plots size against "rank" for our small sample. Except for swapping what was close to the origin with what was far away, this change in approach has merely taken one J-shaped curve and replace by another. If the first was very J-shaped, so too will be the second. How have we gained?

Our gain comes only when we note -- or recall -- that in such situations the product of rank and size often varies far less violently than either alone. In our small example, we have the following:

Rank	Size	Product
1	13	13
2	7	14
3 to 6	1	3 to 6

exhibit 2 of chapter 27x -- Aedes catches

The new approach begun; plot of size vs. rank;
area proportional to number of individual occurrences



Clearly here rank-times-size varies by only a factor of 2 to 4, which is much better than the factor of 13 for raw counts.

* keeping the area principle *

Clearly we are now thinking of plotting rank-times-size vertically. If we are going to keep the area principle, we ought to arrange for the base of the rectangle associated with rank r to be of width proportional to 1/r. While we would like a scale along which the rectangles would rub elbows, no one of us wants to bother to have to add up the reciprocals of the integers.

If we are prepared to make a reasonable approximation, no one of us will have to do this. It turns out that

$$\log(2r+1) - \log(2r-1)$$

is reasonably closely proportional to 1/r, as exhibit 3 shows. Except for r = 1, the failure of proportionality is by much less than three parts in 100 and is thus almost, if not quite, undetectable when we compare areas by looking at rectangles.

(Those who worry about r = 1 can start the interval for r = 1 at any value for 2r-1 between 1.08 and 1.10 to attain better than 3 in 100 precision everywhere. Those who want to adjust two starting values can turn to panel B and get within one part in 100.)

exhibit 3 of chapter 27x -- reciprocals as differences

Relationship between log(2r+1) - log(2r-1) and 1/r

A) UNADJUSTED LOGS

r	log(2r+1) - log(2r-1)	1/r	Ratio
1	.47712	1	.4771
2	.22185	.5	.4437
3	.14613	.33333	.4384
4	.10914	.25	.4366
5	.08715	.2	.4357
6 to 10*	.28083	.64564	.4350
11 to 20*	.29056	.66877	.4345
(large)			(.4343)

Note: logs here are to base 10.

* At these lines we show -- rather than individual values -- the sums of the values belonging in the second and third columns.

B) ADJUSTED LOGS (log 1.1 in place of log 1, log 3.02 in place of log 3)

1	.4383	1	.4383
2	.2192	.5	.4384
3	.14613	.33333	.4384

(and so on)

Note: logs here are to base 10.

We thus preserve the area principle -- in its second form -- if we plot the rth largest "species" as a rectangle with dimensions:

- height proportional to r-times-size.
- base extending from log(2r-1) to log(2r+1).

(We also preserve the area principle if we had added a free start to r both in the product with size and in the log.)

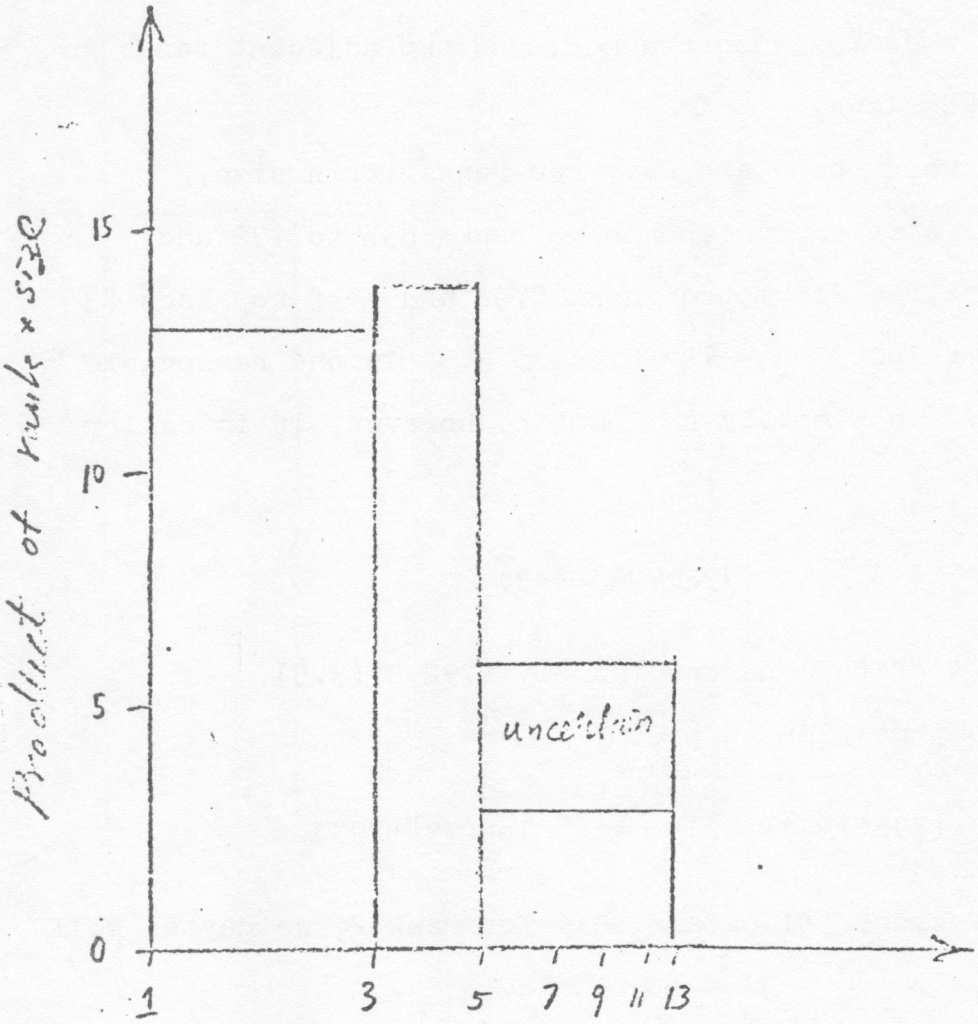
If we now try to make a plot for our small example, we are left with the situation shown in exhibit 4 where the product of rank-times-size for the four species tied at one individual each is shown as somewhere between 1-times-3 and 1-times-6. "What to do in such situations?" is a question best avoided in many circumstances.

27xC. Rank-size-log techniques

We are well on our way to some useful techniques. The form of the area principle appropriate to species has led us to plotting rank-size against log rank. It is time to start modifying our techniques to avoid uncertainties and to gain new features.

exhibit 4 of chapter 27x -- Aedes catches

Naive plot of rank x size against the logarithms of odd integers (trouble with ties)



odd integers (logarithmic scale)

Note: Rank r extends from log(2r-1) to log(2r+1).

* (sum of ranks)×(sum of sizes) *

The first step is to try to plot, not all rectangles but rather only those points about which it is easy to be sure. There are three cases when we can be sure about either a rank or a pair of ranks.

- when there is only one rank for a size.
- when the size changes, and two adjacent ranks go with two sizes.
- when there are only two ranks for a size.

If size 4 corresponds to ranks 637 to 792 and size 3 occurs, we can match ranks "792 and 793" to sizes "3 and 4". Certainly "rank 792.5, size 3.5" is one reasonable combination. To simplify arithmetic, however, it is easier to work with

$$(792+793)(4+3)$$

than with 1/4 of this number, namely (792.5)(3.5).

We thus plan to plot

$$(792+793)(4+3) \quad \text{at} \quad \log(792+793)$$

If size 7341 occurs only for rank 2, we may as well plot

$$(2+2)\cdot(7341+7341) \quad \text{at} \quad \log(2+2)$$

while if size 137 occurs only for ranks 57 and 58 we may as well plot

$$(57+58)\cdot(137+137) \quad \text{at} \quad \log(57+58)$$

The log scale for sum of ranks still begins at 0 = log 1 -- it is now foolish to worry about small adjustments -- and goes to log(2n+1), where n species are represented at least once, so that ranks n and n+1 go with sizes 1 and 0 .

27xD. Examples of RSL use

The calculations and picture for our small example are now as in exhibits 5 and 6.

* back to Josselson *

Turning back to Josselson's Russian word count for The Captain's Daughter, we find calculations -- for selected points only -- and picture as in exhibits 7 and 8. We see a marked constancy of the product between sum-of-ranks 25 and sum-of-ranks 3000. As is so often the case, however, the curve dips at both ends of the scale. The shape of exhibit 8 is thus quite typical.

Since this plot is (nearly enough) area-true to occurrences, we can, for example, judge that about 1/2 of the area lies on each side of sum-of-ranks = 200. Thus the 100 most frequent words account for about 1/2 of all occurrences.

As a general picture of the pattern of fractions of word use, exhibit 8 is most satisfactory. As we shall see shortly, this type of display provides a good general picture for a wide variety of quite and very long-tailed patterns of counts.

exhibit 5 of chapter 27x -- Aedes catches

Calculations for the small example

Rank	Size	Sum of ranks	Sum of sizes	Product
A) SUMS of RANKS and SIZES.				
1	13	2	26	52
		3	20	60
2	7	4	14	56
		5	8	40
3 to 6	1			
		13	1	13
7 on	0			

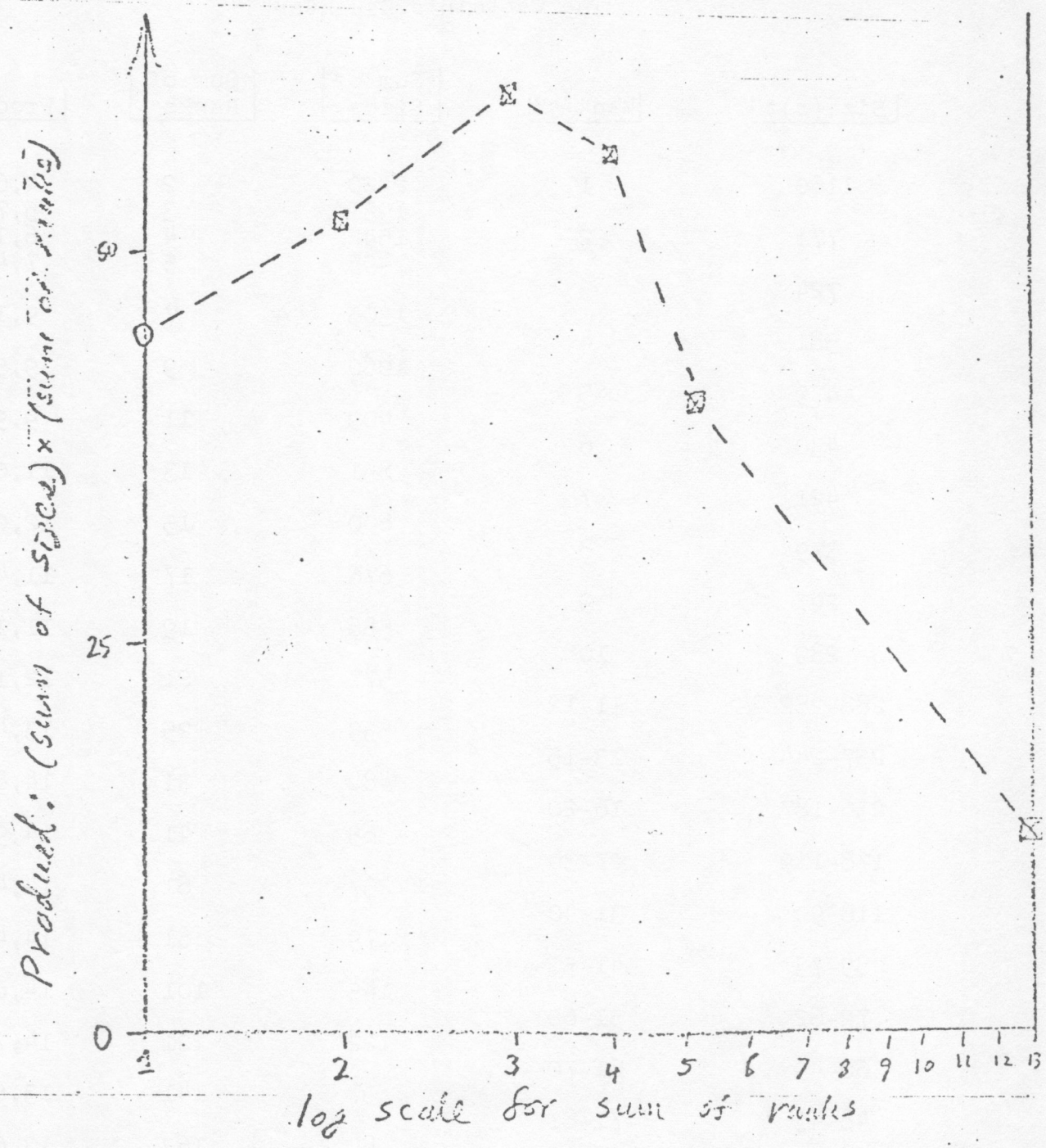
B) STRAIGHT-LINE (in r) EXTRAPOLATION to r = 1, log r = 0.

Sum of ranks	Product
3	60
2	52
1	? (to be found)

$$? = 52 - \frac{60-52}{3-2}(2-1) = 44$$

exhibit 6 of chapter 27x -- Aedes catches

Rank-size-log plot for the small example
(final form; based on exhibit 5)



⊠ Points from exhibit 33 data

O Extrapolated point

exhibit 7 of chapter 27x -- Pushkin's words

Calculations for Josselson's frequency counts for "The Captain's Daughter"

Size(s)*	Rank(s)	Sum of Sizes	Sum of Ranks	Product
1160	1	2320	2	4,640
		1931	3	5,793
771	2	1542	4	6,164
		1495	5	7,475
724	3			
		1305	7	9,135
581	4			
		1060	9	9,540
479	5			
		909	11	9,999
430	6			
		851	13	11,063
421	7			
		800	15	12,000
379	8			
		676	17	11,492
297	9			
		586	19	11,134
289	10			
		577	21	12,111
288-282	11-12			
		539	25	13,475
257-244	13-15			
		480	31	14,880
236-187	16-20			
		365	41	14,965
178-119	21-30			
		237	61	14,457
118-90	31-40			
		179	81	14,499
89-73	41-50			
		145	101	14,645
72-62	51-60			
		122	121	14,762
60-49	61-70			
		97	141	13,677
48-43	71-81			
		84	163	13,692
41-40	82-90			
		78	181	13,766
38-37	91-100			
		73	201	14,673
36-35	101-111			
		69	221	15,249
34-33	112-117			
		65	241	15,665

(Cont'd)

exhibit 7 of chapter 27x (Cont'd)

Size(s)*	Rank(s)	Sum of Sizes	Sum of Ranks	Product
32-31	118-134			
		61	269	16,409
30-29	135-149			
		57	299	16,801
28-27	150-158			
		53	317	16,801
26-25	159-169			
		49	339	16,611
24-23	170-179			
		45	359	16,155
22-21	180-196			
		41	393	16,113
20-19	197-216			
		37	433	16,021
18-17	217-242			
		33	485	16,005
16-15	243-271			
		29	543	15,747
14	272-293			
		27	587	15,849
13	294-313			
		25	627	15,675
12	314-346			
		23	693	15,939
11	347-380			
		21	761	15,981
10	381-419			
		19	839	15,941
9	420-460			
		17	921	15,651
8	461-539			
		15	1079	16,165
7	540-621			
		13	1243	16,159
6	622-735			
		11	1471	16,181
5	736-881			
		9	1763	15,867
4	882-1119			
		7	2239	15,673
3	1120-1552			
		5	3105	15,525
2	1553-2399			
		3	4799	14,397
1	2400-4783			
		1	9567	9,567

* Only sizes (= numbers of occurrences) actually found are mentioned.

** Correcting an obvious misprint.

Source: G. Herdan 1966. The advanced theory of language as choice and chance. (Kommunikation und Kybernetik in einzeldarstellungen. Band 4). Springer, New York. Table 32 at page 97.

exhibit 8 -- Pushkin's words

Rank-size-log rank plot for occurrences of words in Pushkin's The Captain's Daughter

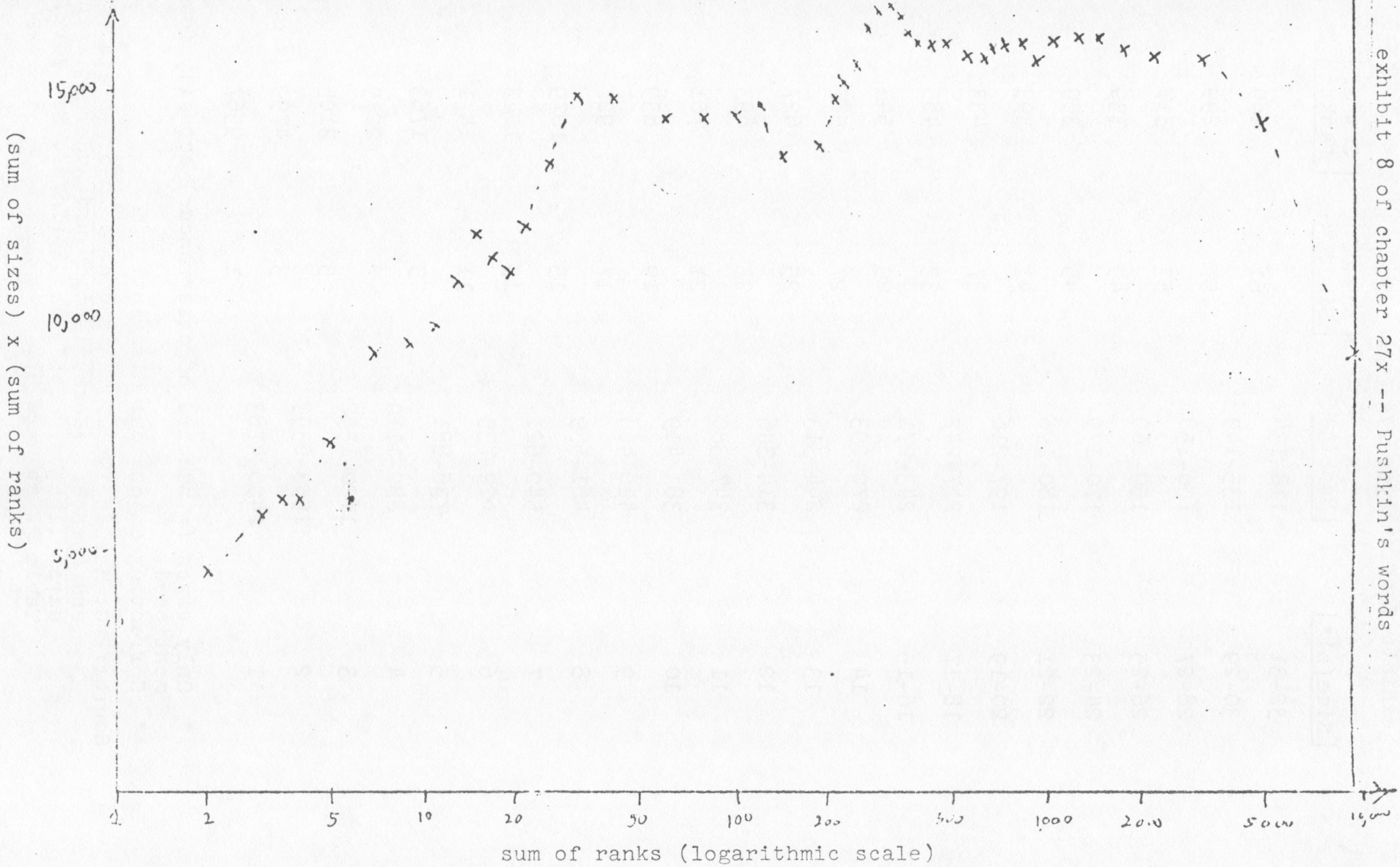


exhibit 8 of chapter 27x -- Pushkin's words

27xE. Raising the ends

Giving a general picture is an important purpose, but there can be others. We might like to look at the pattern of a particular set of counts in more detail, even if we lose the possibility of directly comparing the results of plotting up different sets of data. (We suffer a similar loss, for example, every time we plot residuals from a straight line.) Sometimes we want external comparisons, often we want internal comparisons.

Like many others, our rank-size-log plot is flat in the middle, but droopy at the ends How can we bring up the ends? If we start the ranks, by adding a small constant to the sum of ranks, the rank-1 end will be raised. If we start the sizes, by adding a small constant to the sum of sizes, the size-1 end will be raised.

Let us try this on the Josselson example. Exhibit 9 shows one arithmetic approach, based on choosing a "typical" point in the middle of the array and asking what free starts would be required to make the points at each end match it. Looking at exhibit 8 we see that the unstarted plot has two plateau levels. Panels B and C use a value from the lower plateau. Panel B gives a clear indication of how much start to give the sum of ranks. Panel C gives no such clear indication, tending steadily and vigorously. Panels D and E use

exhibit 9 of chapter 27x -- Pushkin's words

Calculating free starts to make the Josselson data plot "flat"
(see exhibits 7 and 8)

A) LOOKING for a TYPICAL PRODUCT

Exhibit 7 shows 2 plateaus for the (sum of rank)·(sum of sizes) product, one between sum of ranks 31 and sum of ranks 121, the other between sum of ranks 485 and sum of ranks 3105. The products are near 14,700 and 15,700 respectively. We will begin by choosing a point from the first plateau, say 122x121 = 14,762, which has sum of ranks and sum of sizes nearly equal.

B) LARGE SIZES -- to slide rule accuracy, based on exhibit 7

Sum of ranks	Sum of sizes	14,762 over sums	Difference*	Difference after ordering
2	2320	6.35	4.35	3.41
3	1931	7.65	4.65	4.30
4	1542	9.60	5.60	4.31
5	1495	9.85	4.85	4.35
7	1305	11.30	4.30	4.65
9	1060	13.90	4.90	4.80
11	909	16.22	5.22	4.85
13	851	17.31	4.31	4.90
15	800	18.41	3.41	5.22
17	676	21.8	4.80	5.60
				tri/4.66

* Difference = [14,762/(sum of ranks] - [sum of sizes] .

C) SMALL SIZES -- see exhibit 7

Sum of sizes	Sum of ranks	14,762 over same	Difference**
1	9567	1.54	.54
3	4799	3.08	.08
5	3105	4.74	-.26
7	2239	6.60	-.40
9	1763	8.37	-.63
11	1471	10.02	-.98
13	1243	11.83	-1.17
15	1079	13.61	-1.39
17	921	16.00	-1.00
19	839	17.60	-1.40

(Not to be summarized in view of the strong trend.)

** Now [14,762/(sum of ranks)] - [sum of sizes] .

(Cont'd)

exhibit 9 of chapter 27x (Cont'd)

D) SMALL SIZES -- another try; high plateau

		16,801 over same	Difference	ordered differences
1	9567	1.75	.75	.41
3	4799	3.50	.50	.41
5	3105	5.41	.41	.50
7	2239	2.50	.50	.50
9	1763	9.55	.55	.51
11	1471	11.41	.41	.55
13	1243	13.51	.51	.60
15	1079	15.60	.60	.75
17	921	18.22	1.22	
19	839	20.04	1.04	

median .50

E) LARGE SIZES -- a similar try; high plateau

Sum of ranks	Sum of sizes	16,801 over same	Difference	ordered differences
2	2320	7.25	5.25	5.25
3	1931	8.70	5.70	5.70
4	1542	10.89	6.89	5.89
5	1495	11.23	6.23	6.00
7	1305	12.89	5.89	6.23
9	1060	15.82	6.82	6.70
11	909	18.50	7.50	6.82
13	851	19.70	6.70	6.89
15	800	21.00	6.00	7.50
17	616	24.88	7.88	7.88

median 6.46

F) WHAT HAVE WE LEARNED?

The following choices seem possible:

Start for sum of ranks	Start for sum of sizes	Effect for low ranks	Effect for low sizes
4.7	None	Flattened*	Unchanged
4.7	0.5	Flattened*	Flattened**
6.5	0.5	Flattened**	Flattened**

* = near 14,762
** = near 16,801

(Cont'd)

exhibit 9 of chapter 27x (Cont'd)

G) TRIAL of these ALTERNATIVES -- selected points only

(1)	(2)	(3)	(4)			
Sum of ranks		Sum of sizes		Products/1000		
+4.7	+6.5	As is	+0.5	(1)x(3)	(1)x(4)	(2)x(4)
6.7	8.5	2320	2320.5	15.6	15.6	19.7
7.7	9.5	1931	1931.5	14.9	14.9	18.3
8.7	10.5	1542	1542.5	13.4	13.4	16.2
9.7	11.5	1495	1495.5	14.5	14.5	17.2
11.7	13.5	1305	1305.5	15.2	15.2	17.6
13.7	15.5	1060	1060.5	14.5	14.5	16.4
15.7	17.5	909	909.5	14.3	14.3	15.9
17.7	19.5	851	851.5	15.1	15.1	16.6
23.7	25.5	586	586.5	13.9	13.9	15.0
25.7	27.5	577	577.5	14.8	14.8	15.9
35.7	37.5	480	480.5	17.1	17.2	18.0
45.7	47.5	365	365.5	16.7	16.7	17.4
65.7	67.5	237	237.5	15.6	15.6	16.0
105.7	107.5	145	145.5	15.3	15.4	15.6
205.7	207.5	73	73.5	15.0	15.1	15.2
321.7	323.5	53	53.5	17.1	17.3	17.3
547.7	549.5	29	29.5	15.8	16.1	16.1
697.7	699.5	23	23.5	16.1	16.4	16.4
843.7	845.5	19	19.5	16.0	16.4	16.4
925.7	927.5	17	17.5	15.7	16.2	16.2
1083.7	1085.5	15	15.5	16.3	16.8	16.8
1247.7	1249.5	13	13.5	16.3	16.9	16.9
1475.7	1479.5	11	11.5	16.3	17.0	17.0
1767.7	1769.5	9	9.5	15.9	16.8	16.8
2243.7	2245.5	7	9.5	15.7	16.8	16.8
3109.7	3111.5	5	5.5	15.5	17.1	17.1
4803.7	4805.5	3	3.5	14.4	16.8	16.8
9571.7	9573.5	1	1.5	9.6	14.4	14.4

(Cont'd)

exhibit 9 of chapter 27x (Cont'd)

H) TRIAL of +5.6 and +0.7 -- all data from exhibit 7

Sum of ranks +5.6	Sum of sizes +0.7	Product over 1000	Sum of ranks +5.6	Sum of sizes +0.7	Product over 1000
7.6	2320.7	17.6	274.6	61.7	17.0
8.6	1931.7	16.6	304.6	57.7	17.6
9.6	1542.7	14.8	321.6	53.7	17.4
10.6	1495.7	15.9	344.6	49.7	17.2
12.6	1305.7	16.5	364.6	45.7	16.7
14.6	1060.7	15.5	398.6	41.7	16.7
16.6	909.7	15.1	438.6	37.7	16.5
18.6	851.7	15.8	490.6	33.7	16.5
20.6	800.7	16.5	548.6	29.7	16.3
22.6	676.7	15.3	592.6	27.7	16.4
24.6	585.7	14.4	632.6	25.7	16.3
26.6	577.7	15.4	698.6	23.7	16.5
30.6	539.7	16.5	766.6	21.7	16.6
36.6	480.7	17.6	844.6	19.7	16.6
46.6	365.7	17.0	926.6	17.7	16.4
66.6	237.7	15.8	1084.6	15.7	17.1
86.6	179.7	15.6	1248.6	13.7	17.1
106.6	145.7	15.5	1476.6	11.7	17.3
126.6	122.7	15.5	1768.6	9.7	17.2
146.6	97.7	14.3	2244.6	7.7	17.3
168.6	84.7	14.3	3110.6	5.7	17.7
186.6	78.7	14.7	4804.6	3.7	17.8
206.6	73.7	15.2	9572.6	1.7	16.2
226.6	69.7	15.8			
246.6	65.7	16.2			

a value from the higher plateau -- both indicate plausible starts. Panel F reports which combination of starts seem worth trying -- panel G tries them.

Looking hard at panel G suggests using a compromise start for the sum of ranks, say halfway between the two there shown. It also hints that 0.5 may still be a little too small for a start for the sum of sizes. Panel H tries out these indications.

Exhibit 10 shows -- to a substantially enlarged scale -- how panel H works out. At its right end we show the effects of starting by +.5 or +.6 instead of +.7. It would seem that +.6 is probably as good a compromise as we can make. (The behavior at the left end offers no cause for complaint.)

We can summarize Josselson's counts quite well, therefore, by

$$14{,}300 \leq (\text{sum of rank} + 5.6)(\text{sum of sizes} + 0.6) \leq 17{,}700$$

with a remark that there are three or four swings near each limit. This is a pretty good four-number summary for what would otherwise be a difficult-to-manage pattern of counts.

27xF. <u>Broader usefulness</u>

We ought also to try rank-size-log techniques on some of the long-tailed distributions of counts that we earlier found well treated with logs.

exhibit 10

Josselson's count of Pushkin words

Rank-size-log plot with raised ends

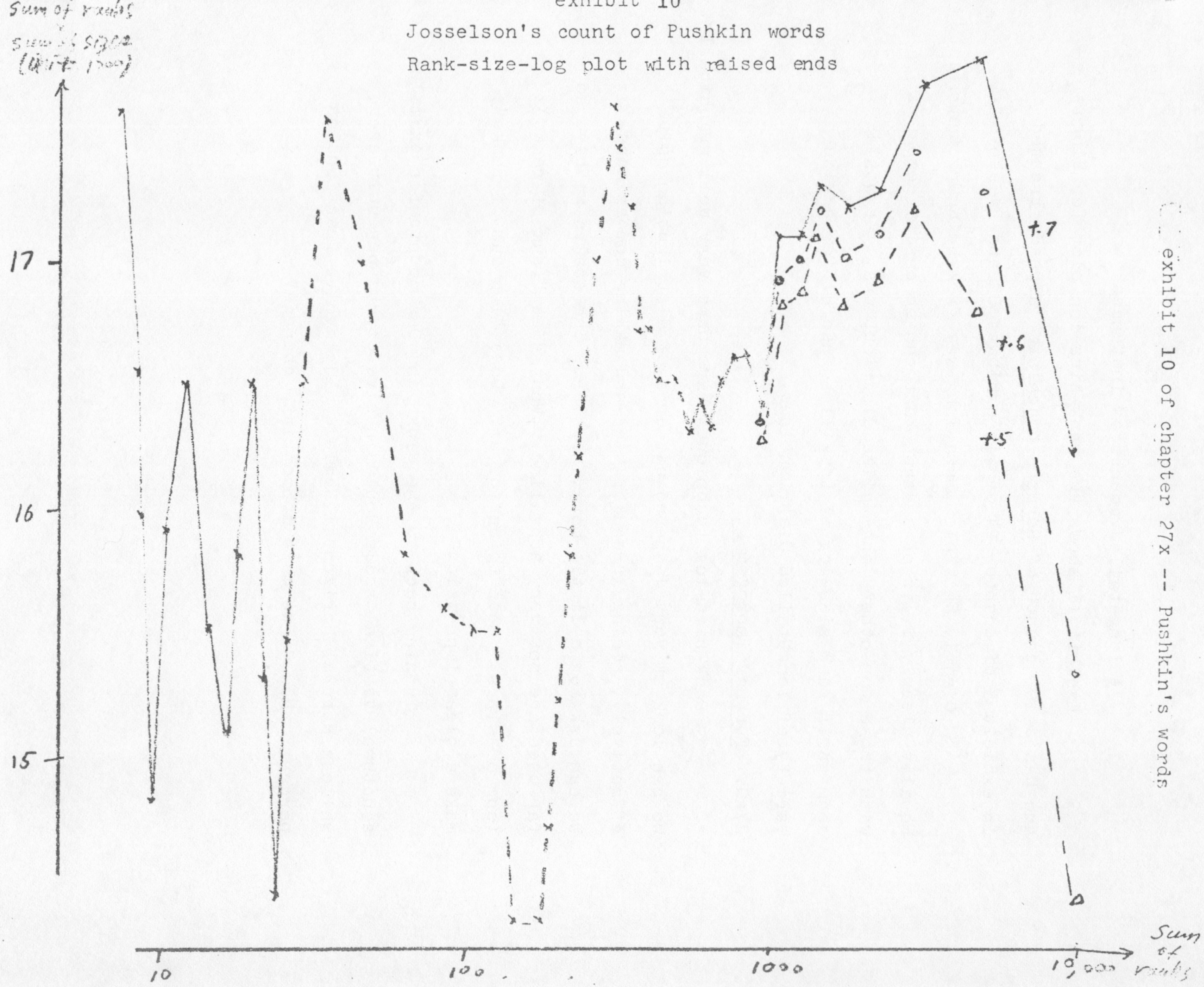


exhibit 10 of chapter 27x -- Pushkin's words

* long-tailed counts of instances *

Exhibit 11 shows the rank-size-log plot for Singh and Das's 100 counts of Bornaya brachiata, already treated in exhibit 5 of chapter 24 and in exhibits 23 and 24 above.

Clearly this plot is a long way from being rectangular. Equally clearly it is of a shape that can be easily compared with those for other long-tailed distributions. It would seem as if rank-size-log plots will be useful for a wide variety of long-tailed distributions.

27xG. Further improvements

Rank-size-log techniques have done very well graphically. We can look at very J-shaped patterns of distribution quite effectively, seeing irregularities and making comparisons. We can ask for a little more, however. The scales we have so far used are not very appealing. What does "(sum of ranks)x (sum of sizes)" mean to any of us? Does "log sum of ranks" make a clear impression?

Turning back to the Singh and Das counts for example, a natural vertical scale would involve counts per unit area, while a natural horizontal scale would involve fractions or percent. How do we make appropriate changes?

exhibit 11 of chapter 27x -- Bornaya infestation

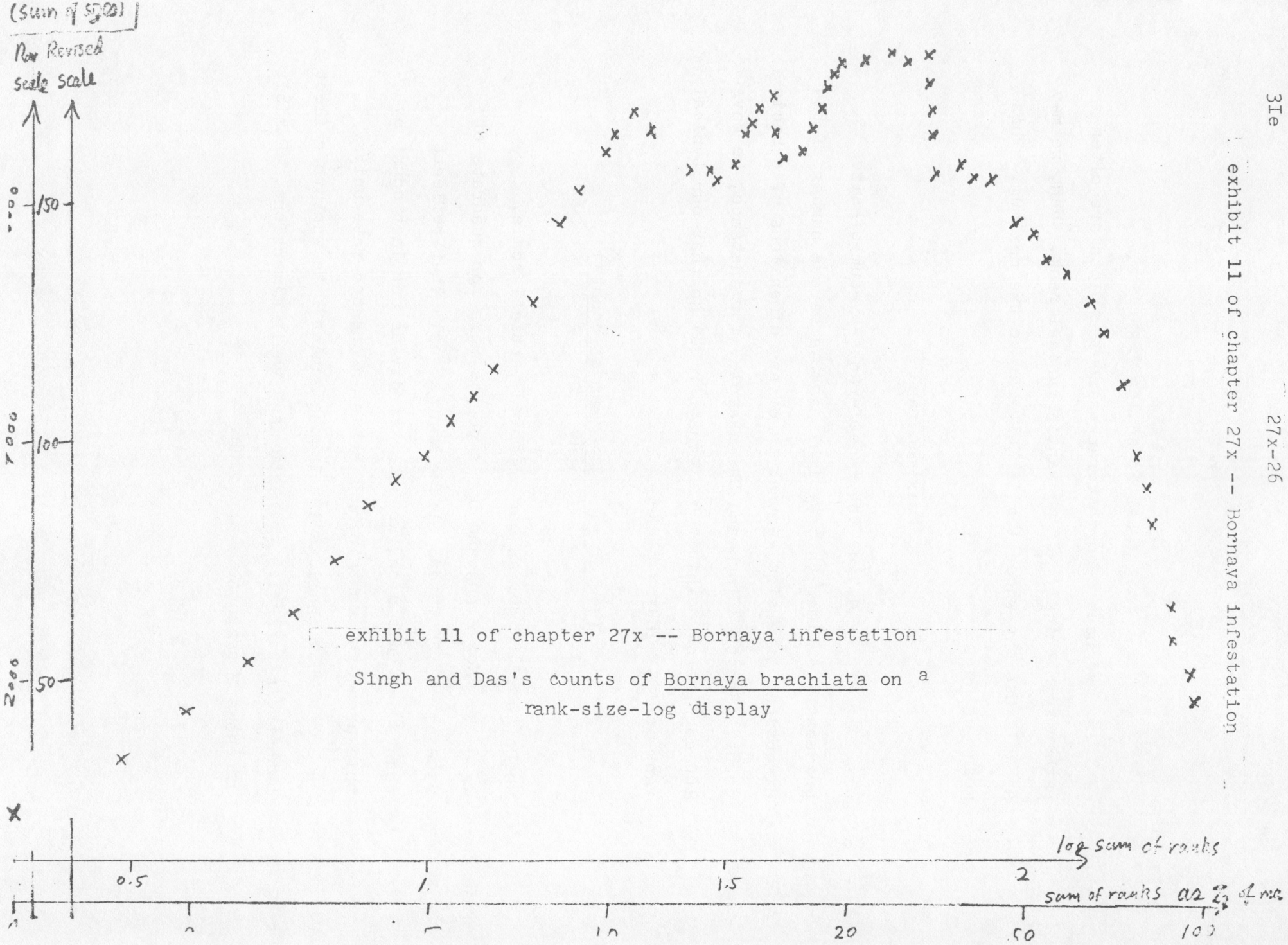


exhibit 11 of chapter 27x -- Bornaya infestation

Singh and Das's counts of Bornaya brachiata on a rank-size-log display

* better scales *

We have seen that rank-size-log plots are often rather symmetrical. If we want a typical rank, then, we may as well use one whose log is half-way to the maximum. This means using

$$\sqrt{\text{maximum rank}}$$

If we divide the sum-of-ranks-by-sum-of-sizes product by 4 times this typical rank we have a number of the dimensions of a size -- that is of the dimensions of a basic count. For the Singh and Das example, for instance, we have 100 counts (see exhibit 5 of chapter 28) so that our typical rank is 10. Thus we take

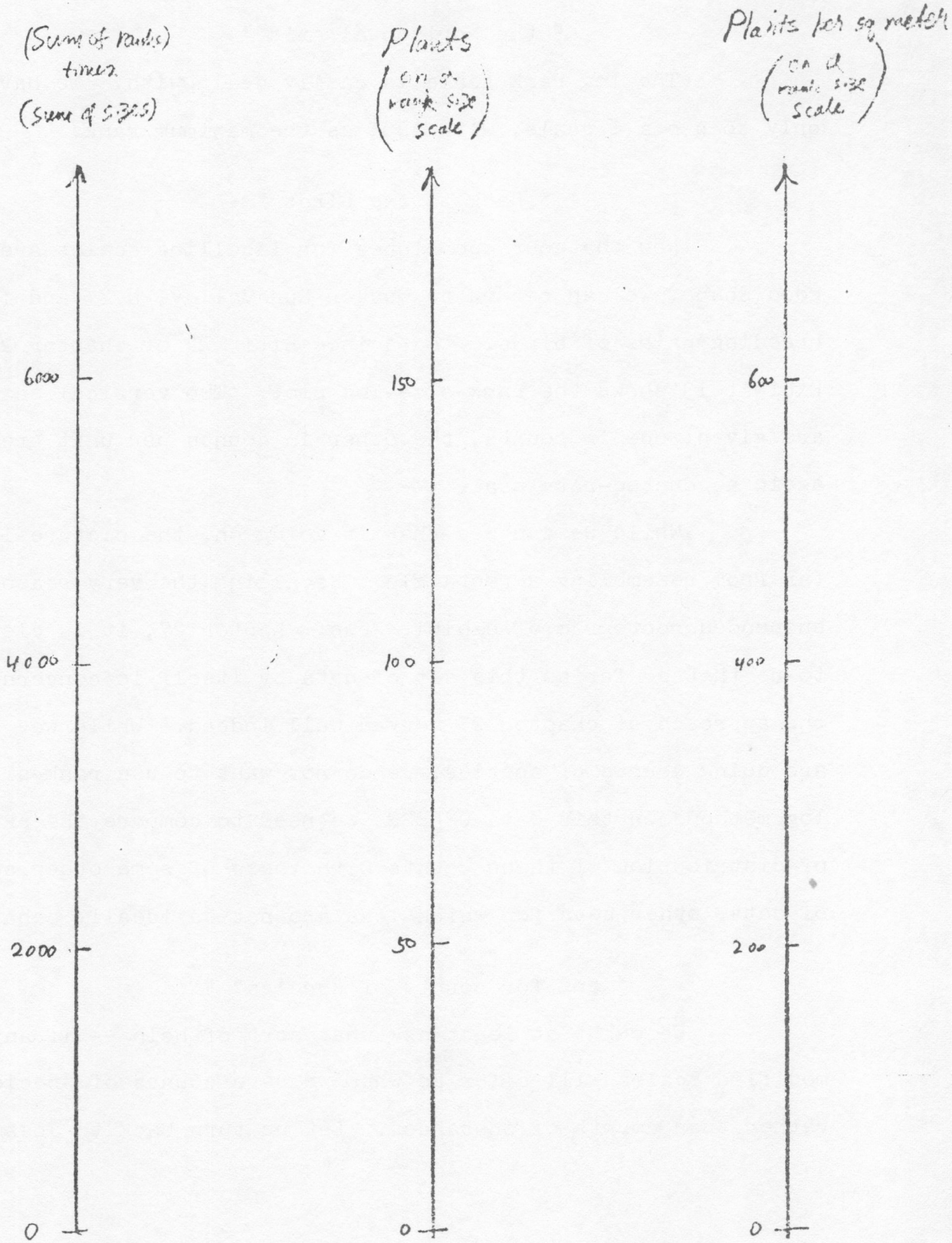
$$\frac{(\text{sum of ranks}) \times (\text{sum of counts})}{40}$$

where $40 = 4 \times 10$, as our improved vertical scale.

Now Singh and Das's counts were for quadrats of $2500\ cm^2 = .25$ meter. If we want a scale that reflects counts per unit area, we ought to divide our last count by this area, thus obtaining a scale in counted-rate units.

How shall we label such scales? We want to reflect both their rank-size character and their dimensions. Exhibit 12 shows a reasonable approach.

Three vertical scales for exhibit 11



* the horizontal axis *

The log rank scale is easily dealt with. We have only to use a % scale, with 100% as the maximum rank.

* back to the birds *

Now that our techniques for labelling scales are in good shape, we can return to Quaker Run Valley, N.Y. and its breeding pairs of birds. (Data in exhibit 24 of chapter 27.) Exhibit 13 shows the rank-size-log plot. Two vertical scales are given; one in counts, the other in counts per unit area, again a counted-rate scale.

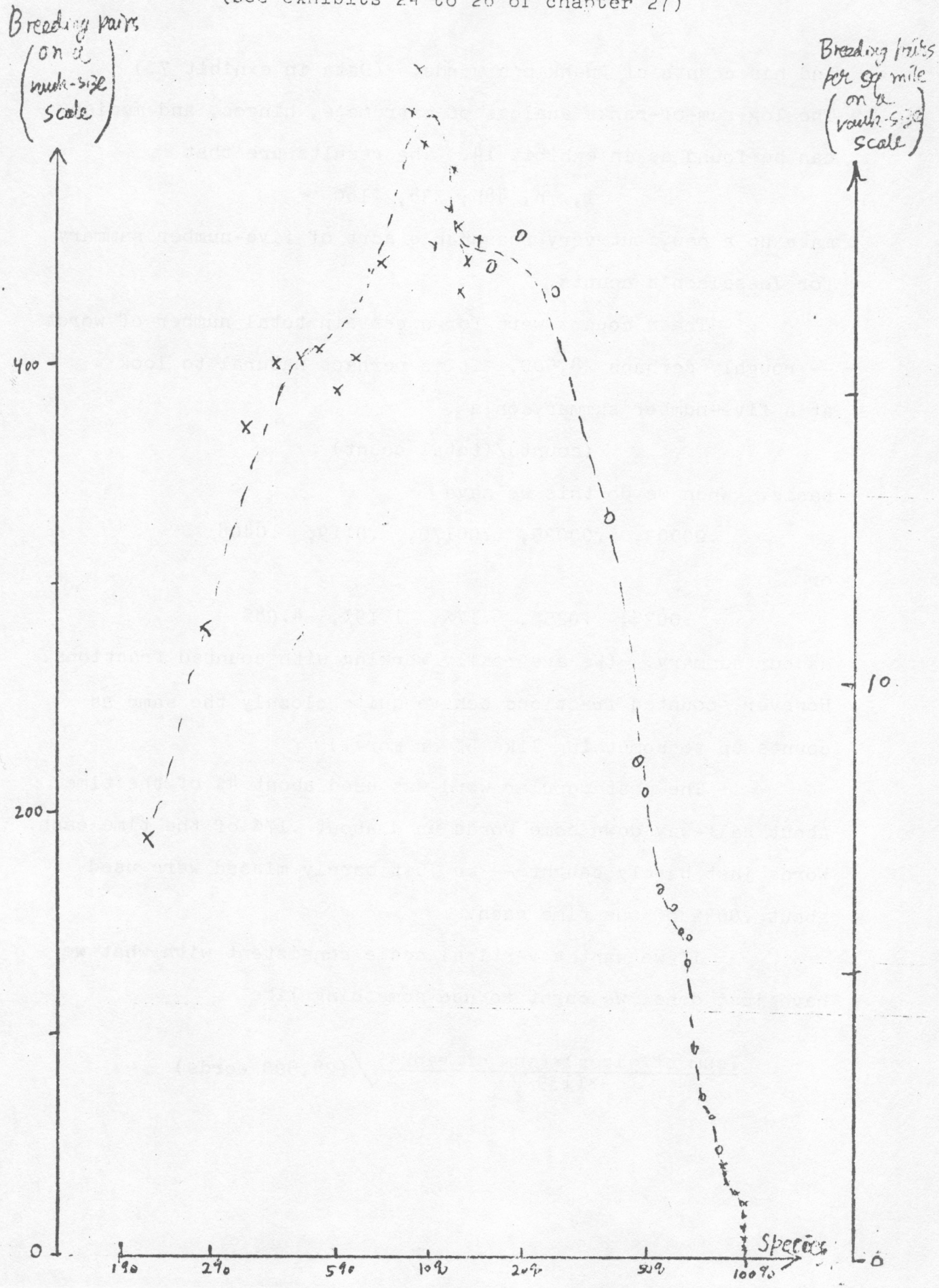
While we can see what is going on, the picture is far from resembling a rectangle. Recalling the very reasonable suspended rootogram of exhibit 26 of chapter 27, it is clear to us that so far as this set of data by itself is concerned, the approach of chapter 27 worked well indeed. While we are doing counts of species, we do not want to use rank-size-log methods on this data UNLESS we need to compare the pattern of-distribution of these counts with those of some other sets of data, other sets for which logs are not so ideally behaved.

* and for counts of species? *

We ought at least ask what sort of help -- if any -- modified scales will offer us when we have counts of species rather than counts of instances. Let us turn back to Josselson

exhibit 13 of chapter 27x -- breeding pairs

The rank-size log plot with improved scales for the birds of Quaker Run Valley, N. Y. (see exhibits 24 to 26 of chapter 27)



and his counts of Pushkin's words. (Data in exhibit 7.) The log-sum-of-ranks analogs of extreme's, hinges, and median can be found as in exhibit 14. The results are that

1, 7h, 48h, 338, 1160

make up a new, but very reasonable sort of five-number summary for Josselson's counts.

These counts were for a certain total number of words -- roughly perhaps 28,500. It is perhaps natural to look at a five-number summary on a

(count)/(total count)

basis. When we do this we have

.00003, .00025, .00170, .0119, .0408

or

.003%, .025%, .17%, 1.19%, 4.08%

as our summary. (We are really working with counted fractions. However, counted fractions behave quite closely the same as counts up to something like 5% or more.)

The most popular word was used about 4% of the time. About half-way down come words used about .17% of the time each. Words just barely caught -- or just barely missed were used about .003% of the time each.

If we want a vertical scale consistent with what we have just done, we ought to use something like

$$\frac{(\text{sum of sizes})\times(\text{sum of ranks})}{4\times(139/2)} \Big/ (28{,}500 \text{ words})$$

exhibit 14 of chapter 27x -- Pushkin's words

Finding the logarithmic analogs of extremes, hinges, and median for Josselson's Russian word frequencies (see exhibits 7 to 10)

A) FIXING the SUMS OF RANKS

	raw	log
Least sum of ranks:	2	0.30
Greatest sum of ranks:	9567	3.98
Mean of logs*:	139	2.14
Again mean of logs**:	16.7	1.22
Again mean of logs***:	1152	3.06

*Alternatively

$$139 = \sqrt{(2)(9567)}$$

**Alternatively

$$16.7 = \sqrt{(2)(139)}$$

***Alternatively

$$1152 = \sqrt{(139)(9567)}$$

Next inward sum of ranks:

17 for 16.7
1079 for 1152

Nearest sum of ranks:

141 for 139

B) The CORRESPONDING SIZES

Sum of ranks	Analog of	Sum of sizes	Size
2	extreme	2320	1160
17	"hinge"	676	338
141	"median"	97	48h
1079	"hinge"	15	7h
9567	extreme	2	1

which, when put in % becomes

$$\frac{(\text{sum of sizes}) \times (\text{sum of ranks})}{79{,}200}$$

Had this scale been used in exhibit 8, the peak value would have been about .21%, indicating the frequency of a middle-use word.

****

27xH. <u>How far have we come?</u>

This chapter has been concerned with very J-shaped patterns of distribution of counts. What very J-shaped means is perhaps best illuminated by something that came up only in section 27xG -- equal spacing in log rank, rather than in rank, as the natural way to find useful way to define 5-number summaries. (Here, as elsewhere in this chapter, rank 1 is the largest count, rank 2 the next, and so on.) There are so many species -- or perhaps, even, instances -- with very small count as to make re-expression of the rank helpful.

For rank-size-log plots we first changed our basic unit from the species to the occurrence, and then moved to plot

$$\frac{(\text{sum of sizes}+?)(\text{sum of ranks}+??)}{4\cdot\sqrt{\text{maximum rank}}}$$

where ? and ?? may represent well-chosen starts, against

$$\log(\text{sum of ranks}+??).$$

Starts and/or denominator are often omitted.

Here each sum is of two terms. Both sums can be of the same number if a given size appears just once. The two sizes can be the same if this size occurs exactly twice. Otherwise the two sizes are different and the corresponding ranks are not only different but adjacent.

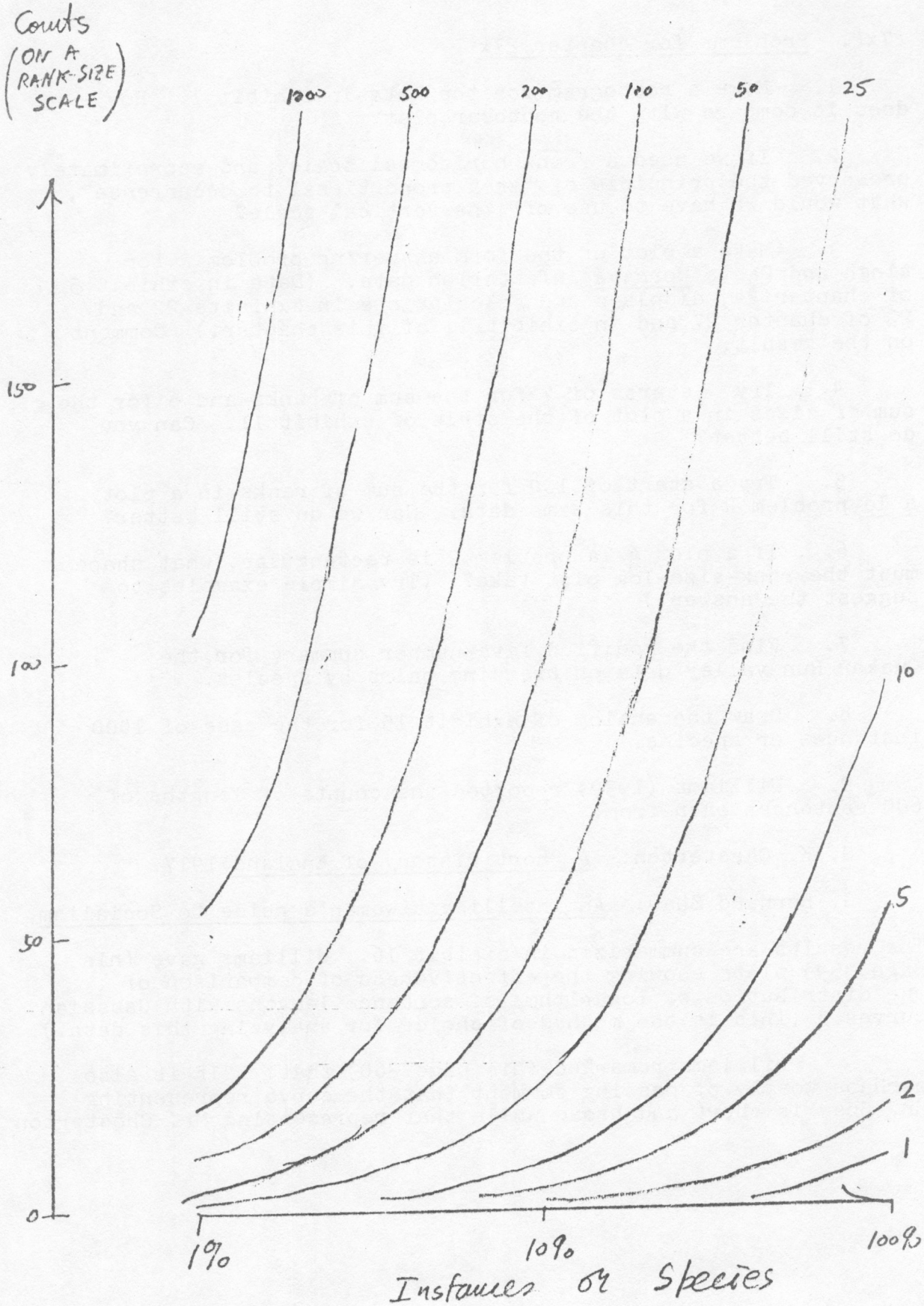
The denominator above was inserted to give the vertical scale the dimensions of a size = a count. Two points are relevant:

- we often want to go further, dividing by an area, a total count, an amount of effort, etc.
- we need to take the size scale with a moderately large grain of salt.

Since the rank-size-log technique is designed to deal with sets of data where conventional plots fail miserably, it could not work satisfactorily without doing something quite different from the conventional. If what it did could not be called -- presumably by its detractors -- "a gross and unwarranted distortion", it could not possibly do as much for us as it does.

Exhibit 15 shows the contours "size=constant" for the case where 100 species or instances are involved. After we recover from the shock, we see that the rank-size scale does apply at $\sqrt{\text{maximum rank}}$ -- here at 10% on the horizontal scale.

Warning about "on a rank-size scale"
Contours of actual counts compared with a "rank-size scale"
for the case of 100 instances or species



****

27xP. Problems for chapter 27x

1. Make a rootogram for the data in exhibit 1. How does it compare with the rootobar plot?

2. If we used a $\sqrt{\text{rank}}$ horizontal scale, and approximately preserved the principle of "area proportional to occurrence", what would we have to use off the vertical scale?

3. Make a plot of the form answering problem 2 for singh and Das's Bornaya infestation data. (Data in exhibit 5 of chapter 24, displays and calculations in exhibits 22 and 23 of chapter 27 and in exhibit 11 of this chapter.) Comment on the result.

4. Try starts of 7 for the sum of ranks and 6 for the sum of sizes in a plot of the style of exhibit 11. Can you do still better?

5. Try a start of 100 for the sum of ranks in a plot á la problem 2 for this same data. Can we do still better?

6. If a plot á la problem 2 is rectangular, what shape must the rank-size-log plot take? (Try simple examples to suggest the answer.)

7. Find the modified five-number summary for the Quaker Run valley data on breeding pairs by species.

8. Draw the analog of exhibit 15 for the case of 1000 instances or species.

9. Williams (1939) reported the counts of lengths of 600 sentences each from:

G. K. Chesterton: A Short History of England 1917.

G. Bernard Shaw: An intelligent woman's guide to Socialism.

The results are summarized in exhibit 16. Williams gave (his page 359) plots showing the effectiveness of comparison of the distribution of logarithms of sentence lengths with Gaussian curves. (This is the method of choice for analyzing this data.)

Williams remarked (his page 360) that: "It is also perhaps worthy of passing comment that the curve representing Mr. Shaw is short and broad while that representing Mr. Chesterton

is tall and slender; which shows how necessary it is to use these curves only for the purpose for which they were originally designed." Make rank-size-log plots for these two distributions in order to see if:

- this (less effective) mode of plotting does manage to contrast these two authors' styles effectively.

- if Williams's remarks apply as well to this plot as to the distribution of logarithms of sentence lengths.

10. In 1962 Struik and Curtis reported the numbers of seeds of various herbs in 1.71 hectares of sugar maple forest. Exhibit 17 has the counts. What expression of these counts seems most reasonable? What happens if a rank-size-log approach is tried? Can we do a good job of raising the ends?

exhibit 16 of chapter 27x -- sentence lengths

Summaries of Williams's Counts of 600 Sentence Lengths

G. K. Chesterton				G. Bernard Shaw			
Length	Count	Length	Count	Length	Count	Length	Count
91	1	29	21	143	1	29	13
76	1	28	22	141	1	28	14
60	1	27	21	115	1	27	13
59	1	26	28	111	1	26	10
58	2	25	19	107	2	25	13
56	2	24	22	99	1	24	13
55	2	23	19	96	3	23	11
53	2	22	26	90	2	22	18
52	2	21	21	87	2	21	12
51	3	20	25	86	1	20	16
50	8	19	24	84	2	19	12
49	2	18	28	83	3	18	22
48	2	17	20	82	1	17	16
47	3	16	16	81	3	16	12
46	4	15	15	80	1	15	21
45	4	14	16	75-79	7	14	16
44	4	13	16	70-74	8	13	14
43	8	12	17	65-67	6	12	6
42	6	11	7	60-64	10	11	9
41	6	10	6	55-59	15	10	15
40	10	9	11	50-54	38	9	16
38-39	13	8	2	45-49	25	8	11
36-37	18	7	5	40-44	28	7	10
34-35	14	6	3	36-39	25	6	10
32-33	29	5	3	34-35	23	5	14
30-31	38			32-33	18	4	3
Total	599			30-31	30	3	2
				Total	600		

Source: C. B. Williams 1939. A note on the statistical analysis of sentence-length as a criterion of literary style. Biometrika 31: 356-361. Dot diagram on page 358.

exhibit 17 of chapter 27x -- seeds by species

Numbers of Seeds Shed in 1.71 Hectares

A) The DATA

Species	Count of seeds
Impatiens pallida	18,400,000
Claytonia virginica	17,050,000
Galium aparine	7,385,000
Aplectum hyemale	5,600,000
Dicentra cucullaria	1,420,000
Trillium recurvatum	594,300
Arisaema triphyllum	592,000
Floerkea prosperpinacoides	317,200
Polygonium pubescens	257,600
Smilacina racemosa	247,000
Sanguinaria canadensis	150,800
Hepatica acatiloba	76,600
Phlox divaricata	47,700
Hydrophyllum virginianum	42,400
Uvularia grandiflora	28,250
Podophyllum peltatum	9,400
Erythronium albidum	560
Caulophyllum thalictroides	960

B) SOURCE

G. J. Struik and J. T. Curtis 1962. Herb distribution in an Acer Saccharum forest. The American Midland Naturalist 68: 285-296.